ERDC/CERL TR-00-12

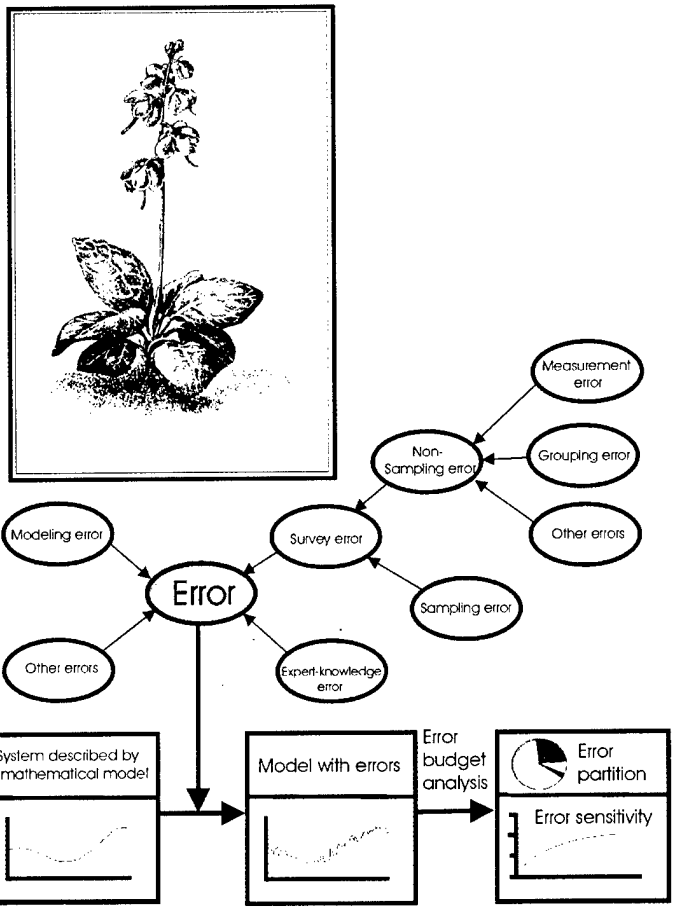
Construction Engineering Research Laboratory

US Army Corps of Engineers®
Engineer Research and Development Center

# Errors In Environmental Assessments: An Error-Budget Model for Plant Populations

by Xiangchi Cao, George Gertner, Bruce MacAllister, and Alan Anderson

April 2000



# Foreword

This study was conducted for the Strategic Environmental Research and Development Program (SERDP) Office under Funding Authorization Document (FAD) 0400-99-8141-08 Work Unit EE9, "Error and Uncertainty for Ecological Modeling and Simulation" (CS-1096). The technical monitor was Dr. Robert Holst, Conservation Program Manager. This study was also conducted for the Office of the Directorate of Environmental Programs, Assistant Chief of Staff (Installation Management) [ACS(IM)], under Project 4A622720A896 "Environmental Quality Technology," Work Unit CN-T09, "Installation Capacity Factors." The technical monitor was Dr. Vic Diersing, DAIM-ED-N.

The work was performed by the Ecological Processes Branch (CN-N) of the Installations Division (CN), Construction Engineering Research Laboratory (CERL). The CERL Principal Investigator was Alan B. Anderson. Part of this work was done by Xiangchi Cao and Dr. George Gertner, of the University of Illinois Department of Natural Resources and Environmental Sciences. Bruce MacAllister worked on the project as an Oakridge Associated Universities postgraduate research fellow. The technical editor was Gloria J. Wienke, Information Technology Laboratory. Steve Hodapp is Chief, CEERD-CN-N and Dr. John T. Bandy is Chief, CEERD-CN. The Acting Director of CERL is Dr. Alan W. Moore.

CERL is an element of the U.S. Army Engineer Research and Development Center (ERDC), U.S. Army Corps of Engineers. The Acting Director of ERDC is Dr. Lewis E. Link and the Commander is COL Robin R. Cababa, EN.

# Contents

# List of Figures and Tables

## Figures



## Tables

# 1 Introduction

## Background

One of the primary missions of the United States Army is to maintain a high state of readiness so it can meet any challenges to national defense. To accomplish this mission, the Army is constantly training soldiers for battle on over 12 million acres of Department of Defense (DoD) lands. The Army is also charged with the stewardship of the lands on which it conducts that training. The Army uses Land Condition Trend Analysis (LCTA) as a means to inventory and monitor natural resources. LCTA was developed by the U.S. Army Construction Engineering Research Laboratory (CERL) under the sponsorship of the U.S. Army Engineering and Housing Support Center (USAEHSC). It uses standardized methods to collect, analyze, and report natural resources data (Diersing, Shaw, and Tazik 1992) as part of the Army's Integrated Training and Management (ITAM) program. An informal review of installation ITAM personnel indicated an interest in estimating plant diversity using LCTA data and modeling changes in plant diversity that result from alternative land uses.

When using a data set like LCTA to model the environment and make management decisions based on that modeling effort, it stands to reason that good, accurate data should be used. The assumption that a data set used for any mathematical or computer modeling is error-free is an underlying premise of theoretical modeling. However, assumptions of error-free data and models usually do not hold true in the real world. Error is a natural property of surveys and modeling and as such, error should be taken into account when developing any type of model.

## Objective

The objective of this project was to develop and test an error-budget model for the population dynamics of plant communities using standard data from the LCTA program at the White Sands Missile Range, New Mexico. Once developed, this error-budget model can in turn be used for a number of other purposes such as data correction, model evaluation, quality control, and management decision-making.

## Approach

The fact that some degree of uncertainty exists in the data used to make management decisions has been recognized by Army installations. Uncertainty and a wide range of probable answers make land management subject to broad interpretation. Natural resource personnel have identified the need for some method to distinguish usable data from unusable data in computer models used to help them in making management decisions. One such pre-existing computer model was in place at White Sands Missile Range (Cao et al. 2000). The authors looked at the data set used for this model for its potential as a test case and chose to use the model as a test case for implementation of a mathematical error-budget model. This error-budget model was developed with input obtained through literature review, professional discussions, and available field data.

## Scope

The error-budget model detailed in this report is designed to improve the use of plant population models. The results of this study are specifically applicable only to the White Sands Missile Range plant population model. By managing for plant communities, DoD has the opportunity to conserve multiple species simultaneously. Plant communities also provide a useful basis on which to understand and manage the natural communities that support military training and other land uses.

Within the context of the larger DoD mission, the use of an error-budget model allows investigators to identify errors in methodology, sampling, data collection, and recording, modeling, and analysis. This process will allow natural resource personnel to make more informed decisions as to what courses of action are appropriate for a given management scenario. Better management of natural resources at the installation level will lead to reduced restrictions on the military mission.

## Mode of Technology Transfer

The information in this report will be provided to Army personnel responsible for assisting with natural resource management issues. The information will also be provided to organizations responsible for developing and refining natural resource conservation methodologies through hard copy reports and through the CERL web site (www.cecer.army.mil).

The error budget for the plant population model included in this report is part of a larger research effort that is developing protocols and tools to account for uncertainty in natural resources modeling efforts and decisionmaking processes. This broader research effort involves developing error budgets for a range of natural resources models as a way of evaluating the uncertainty analysis tools and protocols.

# 2 Error-Budget Modeling

Surveys of plant populations are an integral part of natural resource management. When a survey of a population is completed, the results of the survey are used by decisionmakers to make quantitative statements about the population being studied. This quantitative information helps managers make decisions or perform actions that will affect that population. Errors in these statements can lead to erroneous decisions and actions that have the potential to cause substantial losses to the species the natural resource managers are charged with protecting. Thus, these errors should be carefully studied before committing time and resources to any management project. An error-budget model is used to trace the sources of error and their effects on the quantitative statements that are made using sample data collected in the field. The importance of an error-budget model cannot be overstated when dealing with natural resources.

First, the error-budget model evaluates the quantitative statements made from a survey. Given all the errors, the error-budget model can tell if the statements made are valid or invalid. A statement is valid only if its error is within certain limits. If a statement's error does not fall between the accepted parameters, that statement will usually provide little useful information. Second, the error-budget model can guide survey decisions. Using error sensitivity analysis, all types of error sources can be tested to determine their effects on the final statement. Knowing the effect each source of error has on a statement, we can tailor a survey effort to control those error sources that contribute the most to the final error. This is done based on the sensitivity of the error sources. In this way we may obtain maximum accuracy with minimum cost. Third, an error-budget model provides the information that can be used for error correction. To correct errors, we must first know the sources of the errors. Errors emanating from different sources may require different procedures to correct them. Using error decomposition, we can determine the major causes of the errors. Figure 1 shows the basic components of an error-budget model.

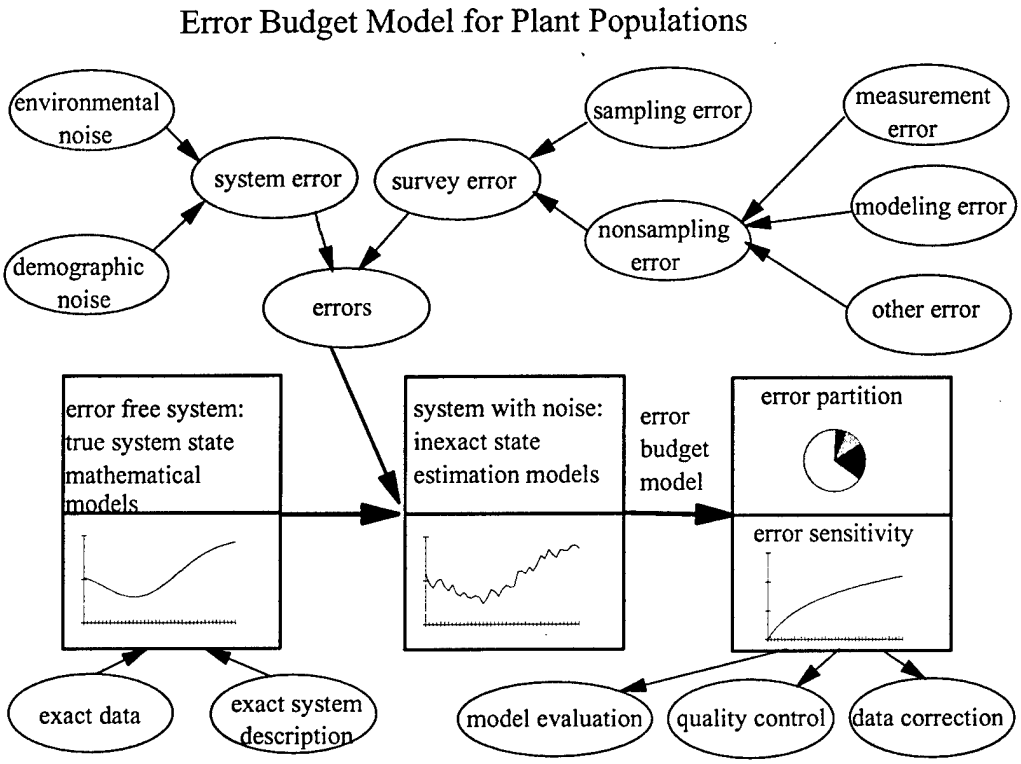



**Figure 1. A conceptual model of an error budget for a plant population model.**

## Error Sources

Errors are classified into two basic categories: system error and survey error. Distinctions between these two errors are based on their sources. System error is a natural character of the system being modeled. It is determined by the system itself. There are two types of system errors: demographic noise and environmental noise (Gotelli 1998). Demographic noise (or within-individual variability) is the variation between individuals who are apparently identical but have different life spans and produce different numbers of offspring. Stochastic models are typically used to investigate the consequences of demographic noise. Environmental noise is so termed because of the fact that changes in the environment vary unpredictably through time. These changes affect individuals in different ways and at different times. The theory of stochastic process can be used to handle both types of system error.

Survey error is the deviation of any survey value from the true value. Survey errors are generally divided into two types: sampling errors and nonsampling errors. Sampling errors inherent in the survey design result from the conscious choice to study a subset rather than the population as a whole. Efforts to control sampling error are grounded in a well-developed theory, as are the formulas and

random selection techniques suitable to a particular problem that falls within the context of the theory. Sampling errors are not the result of mistakes per se, but mistakes in judgment when designing a sample may result in larger errors.

Nonsampling errors encompass all the other things that contribute to survey errors. Nonsampling errors are often thought of as being due entirely to mistakes and deficiencies during the development and execution of the survey procedures. These errors are said to arise from wrongly conceived definitions, imperfections in the tabulation plans, misspecification errors, misclassification errors, and so on. A perfect design would be free of nonsampling errors. The following is a list of some error sources:

1. System error. This error is controlled by the system itself. The survey usually has little to do with it. Choosing appropriate theoretical models is essential to the modeling of the system errors. System errors include environmental noise and demographic noise.
2. Survey error. Survey error consists of sampling error and nonsampling error.
   a. Sampling error. Survey estimates are subject to sampling error because only a subset of the population is measured. The cause of the sampling error is due to the heterogeneity of the population. This error is determined by the population distribution and sampling design.
   b. Nonsampling errors include modeling errors, measurement errors, and other errors.
      i. Modeling errors.
         - Simple models. When simple mathematical models study a complicated population, investigators have only an approximate description of the population. For example, this type of error occurs when a linear model is used to approximate a nonlinear population.
         - Parameterization error. Parameters in the models are usually created by estimation. When estimated parameters are used, the results from the model may be quite different from those that result from theoretical parameters.
         - Projection errors. These errors include prediction error and recursion error. Prediction errors are those errors that occur when we use current model-based information to make a prediction about an unknown future. Recursion error is the error that is compounded by recursive use of models with error.
         - Misspecification errors. These errors occur when the model or model parameters are misspecified.
      ii. Measurement errors. As the name implies, measurement errors refer to the error incurred when the recorded value measured on a study variable

differs from the true value. This error occurs during the data collection stage.

- Instrument error. Tools for recording the values of study variables usually have limited precision. These tools may give inaccurate readings.
- Observer's error. Observers with different backgrounds and training levels will report data with differing levels of accuracy.
- Temporal and spatial errors. These errors occur when the study variable changes with time and place. For example, vegetation has seasonal changes.
- Mistakes or recording errors. Errors of this type occur when researchers make mistakes reading instruments or recording the data. Misclassification of data is also considered a recording error.

iii. Other errors.

- Computation errors. These errors can be avoided with the accuracy of computations made by modern computers.
- Errors due to catastrophe. These errors include lost or destroyed data sampling units.
- Human errors. These errors include typing and editing errors, gaps in knowledge, subjective errors, and so on.

## Error Propagation Method

The following section (pp 11 through 20) is reprinted from *Forest Ecology and Management*, Vol 71; George Gertner, Xiangchi Cao, and Huirong Zhu; "A quality assessment of a Weibull based growth projection system;" pp 235-250; 1995, with permission from Elsevier Science.

In developing an error budget for an inventory/survey system, the first step was to select an appropriate method for determining the effects of errors in the model. The method used was the error propagation method. Error propagation has been used to estimate prediction variances in several models. Gertner (1987, 1988) used this method to determine the prediction variances of STEMS (Belcher, Holdaway, and Brand 1982), a distance-independent growth projection model for the north-central region of the United States. An error propagation method was also used to develop some very simple error budgets for STEMS (Gertner 1990a). Mowrer and Frayer (1986) and Mowrer (1988) used error propagation to estimate prediction variances of several stand-level growth models. In addition, Gertner (1990b) and Gertner and Köhl (1992) used error propagation techniques to assess different inventory systems.

There were a number of reasons why the error propagation method was used for developing the error budget:

- An error budget developed from error propagation is computationally efficient. Using a crude error propagation method to estimate final prediction variance, Gertner (1987, 1988) has shown that results comparable to those of the crude Monte Carlo method can be obtained at only a fraction of the computational cost. Error budgets based on error propagation will have similar computational efficiency.
- Except for testing purposes, high-quality independent data are not necessary for the construction of an error budget based on error propagation. This is true because error propagation methods determine the effect of errors on a model based on the initial properties of that model. Therefore, there is no need to use additional independent data.
- Once appropriate error propagation procedures are incorporated into a multi-component model, error budgets can be generated on-line and the prediction quality can be assessed routinely.
- During a simulation run, the bias and variance of each function in a model can be output regularly. This practice allows for constant monitoring of the accumulation of biases and variances.

To develop error budgets, the error propagation equations for accounting bias, variance, and covariance approximations were extended from those used by Gertner and Mowrer. This was necessary due to the complexities of the combined monitoring-projection system. Extension was also necessitated by the need for very detailed error budgets to conduct the general quality assessments. Since there is concern with the potential problems of model curvature and the resulting biases due to said curvature (Gertner 1991), the assessment was conducted using a second-order Taylor series. The propagation equations were developed to give rise to the biases, variances, and covariances of each component of the model's parameters and predictions through the system. Below is the theoretical development of the error propagation equations used in developing the error budgets.

Assuming an exact function **f** is used to make predictions:

$$\mathbf{Y} = \mathbf{f}(\mathbf{B},\mathbf{X})$$

Where **Y** is a prediction made with the function, $\mathbf{X} = (X_1, X_2, \ldots, X_m)$ is a vector of input variables, and $\mathbf{B} = (b_1, b_2, \ldots, b_m)$ is a vector of known parameters. **X** is usually assumed to be error-free. Now suppose instead of being error-free, the

j-th component of **X**, $\mathbf{X}_j$, has random error $e_j$ (i.e., $X_j = x_j + e_j$) where $e_j$s are independently distributed with mean 0 and variance $V(e_j)$. Then, the predicted **Y** also has error due to the errors of **X**. This error can be estimated by Taylor series expansion.

### *Taylor series expansion method*

Assume any vector function involved has the Taylor series expansion representation up to the second order:

$$(1) \qquad \mathbf{u} = \mathbf{F}(\mathbf{t}) = \mathbf{F}(\mathbf{t}_o) + \frac{\partial \mathbf{F}(\mathbf{t}_o)}{\partial \mathbf{t}^T}(\mathbf{t} - \mathbf{t}_o)$$

$$+ \frac{1}{2}(\mathbf{t} - \mathbf{t}_o)^T \frac{\partial^2 \mathbf{F}(\mathbf{t}_o)}{\partial \mathbf{t}^T \partial \mathbf{t}^{\perp}}(\mathbf{t} - \mathbf{t}_o) + O(|\mathbf{t} - \mathbf{t}_o|^3),$$

where $\mathbf{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_s \end{pmatrix}$, $\mathbf{t} = \begin{pmatrix} t^1 \\ \vdots \\ t_n \end{pmatrix}$, $\mathbf{t}^T$ is the transpose of **t**,

$$\frac{\partial \mathbf{F}(\mathbf{t}_o)}{\partial \mathbf{t}^T} = \begin{pmatrix} \frac{\partial F_1(\mathbf{t}_o)}{\partial t_1} & \cdots & \frac{\partial F_1(\mathbf{t}_o)}{\partial tn} \\ \vdots & \vdots & \vdots \\ \frac{\partial F_s(\mathbf{t}_o)}{\partial t_1} & \cdots & \frac{\partial F_s(\mathbf{t}_o)}{\partial t_n} \end{pmatrix}$$

$$\frac{\partial^2 \mathbf{F}_i(\mathbf{t}_o)}{\partial \mathbf{t} \partial \mathbf{t}^T} = \begin{pmatrix} \frac{\partial^2 f_i(\mathbf{t}_o)}{\partial t_1 \partial t_1} & \cdots & \frac{\partial^2 f_i(\mathbf{t}_o)}{\partial t_1 \partial t_n} \\ \vdots & \vdots & \vdots \\ \frac{\partial 2 f_i(\mathbf{t}_o)}{\partial t_n \partial t_1} & \cdots & \frac{\partial^2 f_i(\mathbf{t}_o)}{\partial t_n \partial t_n} \end{pmatrix}$$

$$(\mathbf{t} - \mathbf{t}_o)^T \frac{\partial^2 \mathbf{F}(\mathbf{t}_o)}{\partial \mathbf{t}^T \partial \mathbf{t}^{\perp}}(\mathbf{t} - \mathbf{t}_o) = \begin{pmatrix} (\mathbf{t} - \mathbf{t}_o)^T \frac{\partial^2 F_1(\mathbf{t}_o)}{\partial \mathbf{t} \partial \mathbf{t}^{\perp}}(\mathbf{t} - \mathbf{t}_o) \\ \vdots \\ (\mathbf{t} - \mathbf{t}_o)^T \frac{\partial^2 F_s(\mathbf{t}_o)}{\partial \mathbf{t}^T \partial \mathbf{t}^{\perp}}(\mathbf{t} - \mathbf{t}_o) \end{pmatrix}$$

If **T** is a random vector and $\varepsilon$ is a random error vector with $E[\varepsilon] = \mathbf{0}$, then

$$(2)\qquad \mathbf{U}-\varepsilon=\mathbf{F}(\mathbf{T})=\mathbf{F}(\mathbf{T}_o)+\frac{\partial \mathbf{F}(\mathbf{T}_o)}{\partial \mathbf{t}^T}(\mathbf{T}-\mathbf{T}_o)$$

$$+\frac{1}{2}(\mathbf{T}-\mathbf{T}_o)^T\frac{\partial^2\mathbf{F}(\mathbf{T}_o)}{\partial \mathbf{t}^T\partial \mathbf{t}^{\perp}}(\mathbf{T}-\mathbf{T}_o)+O_p(|\mathbf{T}-\mathbf{T}_o|^3).$$

If $E[\mathbf{T}]=\mathbf{T}_o$ is true, then from Equation (2) the following can be obtained:

$$(3)\qquad E[\mathbf{U}]=E[\mathbf{F}(\mathbf{T})]\approx \mathbf{F}(\mathbf{T}_o)+\frac{1}{2}E\left[(\mathbf{T}-\mathbf{T}_o)^T\frac{\partial^2\mathbf{F}(\mathbf{T}_o)}{\partial \mathbf{t}^T\partial \mathbf{t}^{\perp}}(\mathbf{T}-\mathbf{T}_o)\right]$$

$$=\mathbf{F}(\mathbf{T}_o)+\frac{1}{2}\begin{pmatrix} Trace[\dfrac{\partial^2\mathbf{F}_1(\mathbf{T}_o)}{\partial \mathbf{t}\partial \mathbf{t}^T}\Sigma_T] \\ \vdots \\ Trace[\dfrac{\partial^2\mathbf{F}_s(\mathbf{T}_o)}{\partial \mathbf{t}^T\partial \mathbf{t}^T}\Sigma_T]\end{pmatrix}.$$

Assuming $\varepsilon$ is independent of $\mathbf{T}$, then

$$(4)\qquad \Sigma_u=\Sigma_\varepsilon+\mathrm{Cov}[\mathbf{F}(\mathbf{T})]$$
$$\approx \Sigma_\varepsilon+E\left[\frac{\partial \mathbf{F}(\mathbf{T}_o)}{\partial \mathbf{t}^T}(\mathbf{T}-\mathbf{T}_o)(\mathbf{T}-\mathbf{T}_o)^T(\frac{\partial \mathbf{F}(\mathbf{T}_o)}{\partial \mathbf{t}^T})^T\right]$$
$$=\Sigma_\varepsilon+\frac{\partial \mathbf{F}(\mathbf{T}_o)}{\partial \mathbf{t}^T}\Sigma_T(\frac{\partial \mathbf{F}(\mathbf{T}_o)}{\partial \mathbf{t}^T})^T$$

Denote $\Sigma_{\mathbf{AB}}=E[(\mathbf{A}-E[\mathbf{A}])(\mathbf{B}-E[\mathbf{B}])^T$ and $\Sigma_A=E[(\mathbf{A}-E[\mathbf{A}])(\mathbf{A}-E[\mathbf{A}])^T]=$ Cov[**A**] for any random vectors **A** and **B**. If there are errors in the variables, for example, if the input vector, instead of **T**, is actually measured as $\tau$, that possesses a bias in **T**: Bias[$\tau$] = E[$\tau$-**T**] = E[$\tau$] - $\mathbf{T}_o$, then the actual explanatory vector of variables,

$$(5)\qquad \nu=\mathbf{F}(\tau)+e=e+\mathbf{F}(\mathbf{T}_o)+\frac{\partial \mathbf{F}(\mathbf{T}_o)}{\partial \mathbf{t}^T}(\tau-\mathbf{T}_o)$$
$$+\frac{1}{2}(\tau-\mathbf{T}_o)^T\frac{\partial^2\mathbf{F}(\mathbf{T}_o)}{\partial \mathbf{t}^T\partial \mathbf{t}^{\perp}}(\tau-\mathbf{T}_o)+O_p(|\tau-\mathbf{T}_o|)^3$$

has a bias in **U**.

$$(6)\qquad \mathrm{Bias}[\nu] = \mathrm{E}[\nu - \mathrm{U}] \approx \frac{\partial \mathbf{F}(\mathbf{T}_o)}{\partial \mathbf{t}^T}\mathrm{Bias}[\tau]$$

$$+\frac{1}{2}\begin{pmatrix} \mathrm{Trace}[\dfrac{\partial^2 \mathbf{F}_1(\mathbf{T}_o)}{\partial \mathbf{t}\partial \mathbf{t}^T}(\Sigma_\tau + \mathrm{Bias}[\tau]\mathrm{Bias}[\tau]^T - \Sigma_T)] \\ \vdots \\ \mathrm{Trace}[\dfrac{\partial^2 \mathbf{F}_s(\mathbf{T}_o)}{\partial \mathbf{t}\partial \mathbf{t}^T}(\Sigma_\tau + \mathrm{Bias}[\tau]\mathrm{Bias}[\tau]^T - \Sigma_T)] \end{pmatrix}.$$

Since

$$(7)\qquad \mathrm{E}[\nu] = \mathrm{E}[\mathbf{F}(\tau)] \approx \mathbf{F}(\mathbf{T}_o) + \frac{\partial \mathbf{F}(\mathbf{T}_o)}{\partial \mathbf{t}^T}\mathrm{E}[\tau - \mathbf{T}_o]$$

$$+\frac{1}{2}\begin{pmatrix} \mathrm{Trace}[\dfrac{\partial^2 \mathbf{F}_1(\mathbf{T}_o)}{\partial \mathbf{t}\partial \mathbf{t}^T}\mathrm{E}[(\tau - \mathbf{T}_o)(\tau - \mathbf{T}_o)^T]] \\ \vdots \\ \mathrm{Trace}[\dfrac{\partial^2 \mathbf{F}_s(\mathbf{T}_o)}{\partial \mathbf{t}\partial \mathbf{t}^T}\mathrm{E}[(\tau - \mathbf{T}_o)(\tau - \mathbf{T}_o)^T]] \end{pmatrix}$$

the covariance matrix becomes

$$(8)\qquad \mathrm{E}\nu = \mathrm{E}_e + \mathrm{Cov}[\mathbf{F}(\tau)]$$

$$\approx \Sigma_e + \mathrm{E}[\frac{\partial \mathbf{F}(\mathbf{T}_o)}{\partial \mathbf{t}^T}(\tau - \mathbf{T}_o - \mathrm{Bias}[\tau])(\tau - \mathbf{T}_o - \mathrm{Bias}[\tau])^T(\frac{\partial \mathbf{F}(\mathbf{T}_o)}{\partial \mathbf{t}^T})^T]$$

$$= \Sigma_e + \frac{\partial \mathbf{F}(\mathbf{T}_o)}{\partial \mathbf{t}^T}\Sigma_\tau(\frac{\partial \mathbf{F}(\mathbf{T}_o)}{\partial \mathbf{t}^T})^T.$$

### *Error evaluation for nonlinear regression system*

Assuming the input-output system is:

$$(9)\qquad \mathbf{Y} = \mathbf{f}(\mathbf{B}, \mathbf{X}) + \varepsilon,$$

where $\mathbf{X} = \begin{pmatrix} X1 \\ \vdots \\ Xn \end{pmatrix}$ and $\mathbf{Y} = \begin{pmatrix} Y1 \\ \vdots \\ Ys \end{pmatrix}$ are respectively the vectors of input and output variables,

$\mathbf{B} = \begin{pmatrix} B1 \\ \vdots \\ Bm \end{pmatrix}$ is the parameter vector, $\varepsilon$ is the random error vector which is independent of $(\mathbf{B}, \mathbf{X})$ and satisfies $\mathrm{E}[\varepsilon] = \mathbf{0}$.

Since the vector function **f(B,X)** is non-linear, it is often difficult to calculate the mean and covariance matrix for the output **Y**, even if it is known that E[**X**] = $\xi$, E[**B**]= $\beta$,

$$\mathrm{Cov}[\mathbf{X},\mathbf{X}] = \mathrm{E}[(\mathbf{X}-\xi)(\mathbf{X}-\xi)^{\mathrm{T}}] = \Sigma_{\mathrm{X}} = \begin{pmatrix} \sigma^2_{\mathrm{X1}} & \cdots, & \sigma_{\mathrm{X1Xn}} \\ \vdots & \vdots & \vdots \\ \sigma_{\mathrm{XnX1}} & \cdots, & \sigma^2_{\mathrm{Xn}} \end{pmatrix},$$

$\mathrm{Cov}[\mathbf{B},\mathbf{B}]=\Sigma_{\mathrm{B}}$, $\mathrm{Cov}[\varepsilon,\varepsilon] = \Sigma_{\varepsilon}$, and $\mathrm{Cov}[\mathbf{B},\mathbf{X}] = \Sigma_{\mathbf{BX}}$. However, a second-order Taylor series expansion can be used to approximate the covariance matrix for the output **Y**. Define, for i = 1, ... , s

$$\mathbf{f}_{\mathrm{i}}^{*} = \mathbf{f}_{\mathrm{i}}(\beta,\xi),$$

$$\mathbf{f}_{\mathrm{iX}}^{*} = \frac{\partial \mathbf{f}_{\mathrm{i}}(\beta,\xi)}{\partial \mathbf{X}} = \begin{pmatrix} \dfrac{\partial \mathbf{f}_{\mathrm{i}}(\beta,\xi)}{\partial X_1} \\ \vdots \\ \dfrac{\partial \mathbf{f}_{\mathrm{i}}(\beta,\xi)}{\partial X_n} \end{pmatrix},$$

$$\mathbf{f}_{\mathrm{iB}}^{*\,\mathrm{T}} = \frac{\partial \mathbf{f}_{\mathrm{i}}(\beta,\xi)}{\partial \mathbf{B}^{\mathrm{T}}} = \begin{pmatrix} \dfrac{\partial \mathrm{f}_{\mathrm{i}}(\beta,\xi)}{\partial B_1} & \cdots, & \dfrac{\partial \mathrm{f}_{\mathrm{i}}(\beta,\xi)}{\partial B_m} \end{pmatrix},$$

$$\mathbf{f}_{\mathrm{iXX}}^{*\,\mathrm{T}} = \frac{\partial^2 \mathbf{f}_{\mathrm{i}}(\beta,\xi)}{\partial \mathbf{X}\partial \mathbf{X}^{\mathrm{T}}} = \begin{pmatrix} \dfrac{\partial^2 \mathbf{f}_{\mathrm{i}}(\beta,\xi)}{\partial X_1 \partial X_1} & \cdots, & \dfrac{\partial^2 \mathbf{f}_{\mathrm{i}}(\beta,\xi)}{\partial X_1 \partial X_n} \\ \vdots & \vdots & \vdots \\ \dfrac{\partial^2 \mathbf{f}_{\mathrm{i}}(\beta,\xi)}{\partial X_n \partial X_1} & \cdots, & \dfrac{\partial^2 \mathbf{f}_{\mathrm{i}}(\beta,\xi)}{\partial X_n \partial X_n} \end{pmatrix},$$

$$\mathbf{f}_{\mathrm{iXB}}^{*\,\mathrm{T}} = \frac{\partial^2 \mathbf{f}_{\mathrm{i}}(\beta,\xi)}{\partial \mathbf{X}\partial \mathbf{B}^{\mathrm{T}}} = \begin{pmatrix} \dfrac{\partial^2 \mathbf{f}_{\mathrm{i}}(\beta,\xi)}{\partial X_1 \partial B_1} & \cdots, & \dfrac{\partial^2 \mathbf{f}_{\mathrm{i}}(\beta,\xi)}{\partial X_1 \partial B_m} \\ \vdots & \vdots & \vdots \\ \dfrac{\partial^2 \mathbf{f}_{\mathrm{i}}(\beta,\xi)}{\partial X_n \partial B_1} & \cdots, & \dfrac{\partial^2 \mathbf{f}_{\mathrm{i}}(\beta,\xi)}{\partial X_n \partial B_m} \end{pmatrix},$$

$$\mathbf{f}_{\mathrm{iBB}}^{*\,\mathrm{T}} = \frac{\partial^2 \mathbf{f}_{\mathrm{i}}(\beta,\xi)}{\partial \mathbf{B}\partial \mathbf{B}^{\mathrm{T}}} = \begin{pmatrix} \dfrac{\partial^2 \mathbf{f}_{\mathrm{i}}(\beta,\xi)}{\partial B_1 \partial B_1} & \cdots, & \dfrac{\partial^2 \mathbf{f}_{\mathrm{i}}(\beta,\xi)}{\partial B_1 \partial B_m} \\ \vdots & \vdots & \vdots \\ \dfrac{\partial^2 \mathbf{f}_{\mathrm{i}}(\beta,\xi)}{\partial B_n \partial B_1} & \cdots, & \dfrac{\partial^2 \mathbf{f}_{\mathrm{i}}(\beta,\xi)}{\partial B_n \partial B_m} \end{pmatrix}$$

and denote:

$$\Delta(\mathbf{B},\mathbf{X}) = \mathrm{Max}_{i,j,k,}\,[|\,X_i - \xi i\,|,|\,B_j - \beta_j\,|,|\,\varepsilon_k\,|].$$

The second order Taylor series expansion of **f** is:

$$(10)\quad f_i(\mathbf{B},\mathbf{X}) = \mathbf{f}_i^* + (\mathbf{X}-\xi)^T f_{iX}^* + f_{iB}^{*T}(\mathbf{B}-\beta) + \frac{1}{2}[(\mathbf{X}-\xi)^T f_{iXX}^{*\,T}(\mathbf{X}-\xi) + 2(\mathbf{X}-\xi)^T f_{iXB}^{*\,T}(\mathbf{B}-\beta) + (\mathbf{B}-\beta)^T \mathbf{f}_{iBB}^{*\,T}(\mathbf{B}-\beta)] + O_p(\Delta^3(\mathbf{B},\mathbf{X})).$$

From Equation (10) the following approximation can be obtained:

$$(11)\quad \eta_i = E[Y_i] = E[\mathbf{f}_i(\mathbf{B},\mathbf{X})] = E[\varepsilon_i] \approx \mathbf{f}^* + E[\mathbf{X}-\xi]^T\,\mathbf{f}_{iX}^* + f_{iB}^{*T} E[\mathbf{B}-\beta] + \frac{1}{2}E[(\mathbf{X}-\xi)^T f_{iXX}^{*\,T}(\mathbf{X}-\xi) + 2(\mathbf{X}-\xi)^T \mathbf{f}_{iXB}^{*\,T}(\mathbf{B}-\beta) + (\mathbf{B}-\beta)^T \mathbf{f}_{iBB}^{*\,T}(\mathbf{B}-\beta)]$$
$$= \mathbf{f}_i^* + \frac{1}{2}(\mathrm{Trace}[\mathbf{f}_{iXX}^{*\,T}\Sigma_X] + 2\,\mathrm{Trace}[\mathbf{f}_{iXB}^{*\,T}\Sigma_{XB}] + \mathrm{Trace}\,[\mathbf{f}_{iBB}^{*\,T}\Sigma_B]).$$

Similarly,

$$(12)\quad \sigma_{Y_iY_j} = \mathrm{Cov}[Y_i,Y_j] = E[Y_i,Y_j] - \eta_i\eta_j \approx \sigma_{\varepsilon_i\varepsilon_j} + \mathbf{f}_{iX}^{*\,T}\Sigma_X\,\mathbf{f}_{jX}^* + \mathbf{f}_{iB}^{*\,T}\Sigma_{BX}\,\mathbf{f}_{jX}^* + \mathbf{f}_{jB}^{*\,T}\Sigma_{BX}\,\mathbf{f}_{iX}^* + \mathbf{f}_{iB}^{*\,T}\Sigma_B\,\mathbf{f}_{jB}^*.$$

In the calculations, all terms of the order $E[\Delta^3(B,X)]$ or higher were omitted from the assessment. Also, since $E[\Delta^4(B,X)] - E[\Delta^2(B,X)]^2 \geq 0$, the term involving products of traces of the two covariance matrices, which has the order of $E[\Delta^2(B,X)]^2 \leq E[\Delta^4(B,X)]$, were also omitted.

Because there can be errors in the actual input x and estimated parameter b, with E[x] = ξ + Bias[x], E[b] =β + Bias[b], Cov[x,x] = $\Sigma_X$, Cov[b,b] = $\Sigma_b$, and Cov[b,x] = $\Sigma_{bx}$, the actual output should be:

(13) $\mathbf{y}=\mathbf{f}(\mathbf{b},\mathbf{x}) + \varepsilon$

Substituting **(B,X)** with **(b,x)** into Equation (2) and taking the expectation, the following is obtained (as above, omit all terms of the order $E[\Delta^3]$ where Δ = Max[Δ(**B,X**), Δ(**b,x**)]):

$$(14)\quad E[y_i] = E[\mathbf{f}_i(\mathbf{b},\mathbf{x})] + E[\varepsilon_i] \approx \mathbf{f}_i^* + Bias[\mathbf{x}]^T \mathbf{f}_{iX}^* + \mathbf{f}_{iB}^{*T} Bias[\mathbf{b}]$$

$$+\frac{1}{2}(\text{Trace}[\mathbf{f}_{iXX}^{*T}(\Sigma_x + \text{Bias}[\mathbf{x}]\,\text{Bias}[\mathbf{x}]^T)]$$

$$+2\,\text{Trace}[\mathbf{f}_{iXB}^{*T}(\Sigma_{bx} + \text{Bias}[\mathbf{b}]\,\text{Bias}[\mathbf{x}]^T)]$$

$$+\text{Trace}[\mathbf{f}_{iBB}^{*T}(\Sigma_b + \text{Bias}[\mathbf{b}]\,\text{Bias}[\mathbf{b}]^T)],$$

$$(15)\quad \text{Bias}[y_i] = E[y_i] - \eta_i \approx \text{Bias}[\mathbf{x}]^T \mathbf{f}_{iX}^* + \mathbf{f}_{iB}^{*T}\ \text{Bias}[\mathbf{b}]$$

$$+\frac{1}{2}(\text{Trace}[\mathbf{f}_{iXX}^{*T}(\Sigma_x + \text{Bias}[\mathbf{x}]\,\text{Bias}[\mathbf{x}]^T - \Sigma_x)]$$

$$+2\,\text{Trace}[\mathbf{f}_{iXB}^{*T}(\Sigma_{bx} + \text{Bias}[\mathbf{b}]\,\text{Bias}[\mathbf{x}]^T - \Sigma_{BX})]$$

$$+\text{Trace}[\mathbf{f}_{iBB}^{*T}(\Sigma_b + \text{Bias}[\mathbf{b}]\,\text{Bias}^T - \Sigma_B)]).$$

The covariance between $y_i$ and $y_j$ is (note that $E[\Delta]^2 \le E[\Delta^2]$ and $E[\Delta^2]^2 \le E[\Delta]E[\Delta^3]$ so $E[\Delta]\,E[\Delta^2] \le E[\Delta^3]$)

$$(16)\quad \sigma_{yiyj} = E[y_i, y_j] - E[y_i]\,E[y_j]$$
$$\approx \sigma_{\varepsilon i\varepsilon j} + \mathbf{f}_{iX}^{*T}\Sigma_x\,\mathbf{f}_{jX}^* + \mathbf{f}_{iB}^{*T}\Sigma_{bx}\,\mathbf{f}_{jX}^* + \mathbf{f}_{jB}^{*T}\Sigma_{bx}\,\mathbf{f}_{iX}^* + \mathbf{f}_{iB}^{*T}\Sigma_b\,\mathbf{f}_{jB}^*$$

In an iterating system like the inventory system, the output serves as the input for the next year. To evaluate the transition error, the following covariance between **b** and **y** is needed, $\Sigma_{by} = \text{Cov}[\mathbf{b},\mathbf{y}]$, which is calculated from:

$$(17)\quad \text{Cov}[\mathbf{b}, y_j] = E[\mathbf{b}y_j] - E[\mathbf{b}]\,E[y_j] \approx \Sigma_{bx}\,\mathbf{f}_{jX}^* + \Sigma_b \mathbf{f}_{jB}^*\,.$$

### *Error transition in iterative system*

In the development of the iterative system, it was assumed that the models were properly specified and calibrated, such that the Bias [**b**] = **0** and E [**b**] = $\beta$. At the beginning, it was assumed that there were only random measurement errors in the initial input $\mathbf{x}^{(0)} = \mathbf{X}^{(0)}$, i.e., Bias[$\mathbf{x}^{(0)}$]=0 and E[$\mathbf{x}^{(0)}$] = $\xi^{(0)}$. Then for the initial input $x^{(0)}$ and theoretical parameter $\mathbf{B} = \beta$ (non-random constant), the output is:

(18) $\mathbf{X}^{(1)} = \mathbf{f}(\beta, \mathbf{x}^{(0)}) + \varepsilon^{(1)}$

and for the same initial input but estimated parameter **b**, the output becomes:

(19) $\mathbf{x}^{(1)} = \mathbf{f}(\mathbf{b}, \mathbf{x}^{(0)}) + \varepsilon^{(1)}$

Assuming the data set used to estimate the parameter is independent of $\mathbf{x}^{(0)}$ and $\mathrm{Cov}[\mathbf{b}, \mathbf{x}^{(0)}] = \Sigma_{bx}^{(0)} = 0$, the bias is as follows:

(20) $\mathrm{Bias}[x_i^{(1)}] = \mathrm{E}[x_i^{(1)} - X_i^{(1)}] \approx \frac{1}{2}\mathrm{Trace}[\mathbf{f}_{iBB}^{(0)^T} \Sigma_b]$.

It can be seen that the bias in the output is created in a one-step transition. For a well fitted model, it can be expected that $f_{iBB}^{(0)^T} = \frac{\partial^2 \mathbf{f}_i(\beta, \xi^{(0)})}{\partial \mathbf{B} \partial \mathbf{B}^T}$ or $\Sigma_b$ is sufficiently small, so the bias can be negligible. But for the model with large $\Sigma_b$, it is necessary to consider the effects of the bias on the future projections since it will use the current output as the next input. The k-th actual output is:

(21) $\mathbf{x}^{(k+1)} = \mathbf{f}(\mathbf{b}, \mathbf{x}^{(k)}) + \varepsilon^{(k+1)}$.

The theoretical output is:

(22) $\mathbf{X}^{(k+1)} = \mathbf{f}(\beta, \mathbf{x}^{(k)}) + \varepsilon^{(k+1)}$

The bias is approximately equal to:

$$(23)\quad \mathrm{Bias}[x_i^{(k+1)}] = E[x_i^{(k+1)}] - E[X_i^{(k+1)}] \approx \mathrm{Bias}[\mathbf{x}^{(k)}]^T \mathbf{f}_{iX}^{(k)} + \frac{1}{2}\mathrm{Trace}[\mathbf{f}_{iXX^T}^{k}(\Sigma_X^{(k)} + \mathrm{Bias}[x^{(k)}]\,\mathrm{Bias}[\mathbf{x}^{(k)}]^T - \Sigma_X^{(k)})] + \mathrm{Trace}[\mathbf{f}_{iXB^T}^{(k)}\Sigma_{bx}^{(k)} + \mathbf{f}_{iBB^T}^{(k)}\Sigma_b]$$

where $\Sigma_x^{(k)}$, $\Sigma_X^{(k)}$, $\Sigma_{bx}^{(k)}$ are calculated through

$$(24)\quad \sigma_{ij}^{(k)} = \mathrm{Cov}[x_i^{(k)} x_j^{(k)}] \approx \sigma_{\varepsilon_i\varepsilon_j}^{(k)} + f_{iX^T}^{(k-1)} \Sigma_x^{(k-1)} f_{jX}^{(k-1)} + f_{iB^T}^{(k-1)} \Sigma_{bx}^{(k-1)} f_{jX}^{(k-1)} + f_{jB^T}^{(k-1)} \Sigma_{bx}^{(k-1)} f_{iX}^{(k-1)} + f_{iB^T}^{(k-1)} \Sigma_b f_{jB}^{(k-1)},$$

$$(25)\quad \sigma_{0ij}^{(k)} = \mathrm{Cov}[X_i^{(k)} X_j^{(k)}] \approx \sigma_{\varepsilon_i\varepsilon_j}^{(k)} + f_{iX^T}^{(k-1)} \Sigma_X^{(k-1)} f_{jX}^{(k-1)} + f_{iB^T}^{(k-1)} \Sigma_b f_{jB}^{(k-1)},$$

and

$$(26)\quad \mathrm{Cov}[b, x_j^{(k)}] \approx \Sigma_{bx}^{(k-1)} f_{jX}^{(k-1)} + \Sigma_b f_{jB}^{(k-1)}.$$

At the first step, since $\Sigma_{bx}^{(0)} = 0$, these terms are simplified as

$$(27)\quad \sigma_{ij}^{(1)} \approx \sigma_{\varepsilon_i\varepsilon_j}^{(1)} + f_{iX^T}^{(0)} \Sigma_X^{(0)} f_{jX}^{(0)} + f_{iB^T}^{(0)} \Sigma_b f_{jB}^{(0)} \approx \sigma_{0ij}^{(1)},$$

$$(28)\quad \mathrm{Cov}[b, x_j^{(1)}] \approx \Sigma_b f_{jB}^{(0)}.$$

Further steps are deduced by the iterate algorithms. In this way the error increase can be approximated in each step.

## Misclassification

Situations in which discrete variables are measured with error are called misclassifications. Classification is the process of dividing objects or items into mutually exclusive groups, such that the members of each group are as "close" as

possible to one another, and different groups are as "far" as possible from one another. The distance is measured with respect to specific variable(s) or properties you are trying to predict. In the visible world, objects are characterized by their properties. These properties are usually measured by discrete variables or categorized continuous variables. Based on their measures, objects are classified into classes. In an error-free world, every object belongs to its right class. In the real world, however, objects may be measured with error and placed into wrong classes. These objects are considered to be misclassified. Misclassification does not change the total number of objects or items, but it changes the distribution of objects among the classes. For example, let $O = (o_1, \ldots, o_{10})$ be the set of 10 objects, $P = (p_1, \ldots, p_k)$ be the k properties of each object in $O$, and $C = (c_1, c_2, c_3)$ be the set of 3 classes. Without error, the class set is $C = \{ \{ o_1, o_5\}, \{ o_2, o_3, o_9, o_{10}\}, \{ o_4, o_6, o_7, o_8\} \}$. When the objects are measured with error, we may have the class set $C = \{ \{ o_5, o_7, o_{10}\}, \{ o_1, o_2, o_3 \}, \{ o_4, o_6, o_8, o_9\} \}$.

There are many causes of misclassification, such as inaccurate measurements of objects, incomplete information about the objects, or human mistakes. The results of a classification are used to make statements about the objects being studied. They can also provide information for decisionmaking. Misclassification errors can lead to an incorrect decision, thereby causing substantial losses. Thus, classification errors should be carefully studied.

Errors in measurement not only cause larger variance, they also may produce bias in the results (Gertner, Cao, and Zhu 1995). Methods of dealing with measurement errors have been proposed and successfully used in applications like error propagation (Gelb et al. 1974; Gertner, Cao, and Zhu 1995) and error approximation (Gertner 1987). These methods however, cannot be applied in instances of misclassification because of its property as a closed system. The problem of misclassification has been considered from the different viewpoint by many investigators. To adjust for misclassification, Tenebein (1979) proposed a double sampling scheme for binomial data. Chen (1979, 1989) gave a review of methods for misclassified categorical data and the maximum likelihood estimation for loglinear models. Geng (1989), York (1992), and Viana (1994) applied Bayesian estimation methods to the problem of misclassification and incomplete data. In this study, we will discuss two approaches to modeling misclassification: likelihood function methods and Bayesian estimation methods. We will apply these methods to the estimation of biodiversity with misclassifications.

### *Likelihood function method*

In some systems, objects are distributed in theoretic patterns. The distribution of objects in these systems can be described precisely in a mathematical fashion.

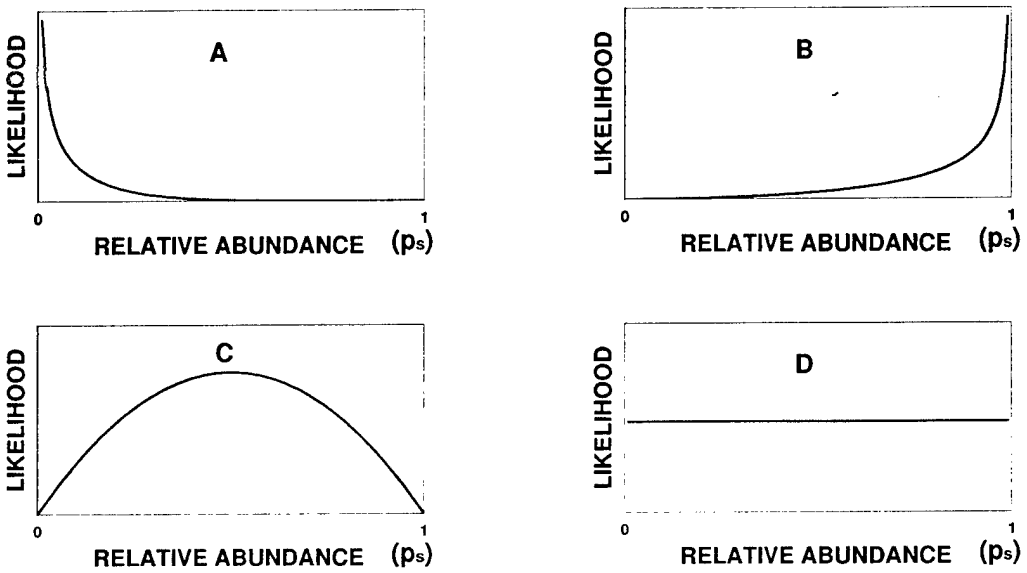
Also, when data or information are not sufficient to make statistical conclusions about the distribution of the objects studied, assumptions of theoretic distributions must be made. Among all the theoretic distribution functions, we find that the beta function has the highest flexibility to model a wide variety of distribution types.

Magnussen and Boyle (1995) use a beta function as an *a-priori* likelihood function to represent the most probable species abundance distributions (MOPSAD). Shannon and Simpson indices are calculated by using MOPSAD. The method of using MOPSAD considers the variations due to sampling. Based on this model, we propose a similar approach for estimating diversity indices with misclassifcation.

Suppose we have the following beta function, (also called a likelihood function) as a species abundance distribution.

$$(29) \quad L(p_s|\alpha,\beta) = p_s^{\alpha-1} \times (1-p_s)^{\beta-1} / B(\alpha,\beta)$$

$L(p_s \mid \alpha, \beta)$ is the likelihood of species s having a relative $p_s$. $\alpha$ and $\beta$ are two parameters of the beta function. $B(\alpha, \beta)$ is the complete beta function of $\Gamma(\alpha)$ $\Gamma(\beta)/\Gamma(\alpha+\beta)$. With different combinations of the $\alpha$ and $\beta$ values, the likelihood function $L(p_s)$ gives different types of curves. Figure 2 displays four typical types of MOPSAD *priori* for plant communities, which are called inverse J-shaped, J-shaped, bell shaped, and flat *priori* by Magnussen and Boyle (1995). The inverse J-shaped distribution usually represents communities in which there are a lot of rare species and very few dominating species. In contrast, the J-shaped curves illustrate a situation in which the communities are dominated by a few principal species. It is unlikely one would find many rare species in these instances. The bell-shaped species distribution appears in many Montane temperate forests. In these forests the two extremes (dominated entirely by rare species and dominated by only a few common species) are rare. The flat *priori* distribution represents an even distribution of species.




(a) Inverse J-shaped $L(p_s|0.5, 3.0)$, (b) J-shaped $L(p_s|3.0, 0.5)$, (c) bell-shaped $L(p_s|2.0, 2.0)$, and (d) a flat *priori* $L(p_s|1.0, 1.0)$.

**Figure 2. Four typical types of species abundance distributions.**

In this study, we use a Shannon index as an example of the error analysis of plant diversity. A Shannon index crystallizes both species richness and species evenness into a single number (Shannon and Weaver 1949). The value of the Shannon index is determined by both the numbers of species and species distributions. The following is the formula for the Shannon index.

$$(30) \quad H = -\sum_{i=1}^{s} p_i \times \log(p_i),$$

H is the Shannon index, $p_i$ is the relative abundance of species i, and s is the total number of species.

The expected Shannon index for a plant community with a beta distribution representing MOPSAD is found by summing all possible relative species abundance's ($0 < p_s \leq 1$) with each summoned given a weight equal to its likelihood $L(p_s)$. Magnussen and Boyle (1995) give the conditional expectation of the Shannon index.

$$(31) \qquad E(H \mid \alpha,\beta) = \frac{\int_0^1 -p_s \times \log(p_s) \times L(p_s) dp_s}{\int_0^1 \times L(p_s) dp_s}$$

$$= \frac{\int_0^1 -p_s^{\alpha} \times (1-p_s)^{\beta-1} \times \log(p_s) dp_s}{B(\alpha,\beta) \times \alpha/(\alpha+\beta)}$$

$\int_0^1 ps \times L(ps) dps = \alpha/(\alpha+\beta)$ is the mean species abundance for the chosen MOPSAD and log is the natural log function.

Based on Equation (31), we give the following the conditional variance of $E(H \mid \alpha, \beta)$:

$$(32) \qquad Var(H \mid \alpha,\beta) = \int_0^1 [(-p_s \times \log(p_s) \times \frac{\alpha}{\alpha+\beta} - E(H \mid \alpha,\beta)]^2 \times L(p_s) dp_s$$

Without misclassification error, Equations (31) and (32) give the expectation and variance of Shannon index (Equation 30). Figure 3 shows the expected Shannon index with α and β ranged from 0.5 to 10.

Misclassification can occur for many reasons in a plant species survey. The major sources of species misclassification are incorrectly identifying species, recording in the wrong catalogue, miscoding species, or using poor quality specimen. Incorrectly identifying species is related to weather, season, and human background on plant study. With the guidance of experts, this error can be very small. Recording errors occur very often among species with similar codes such

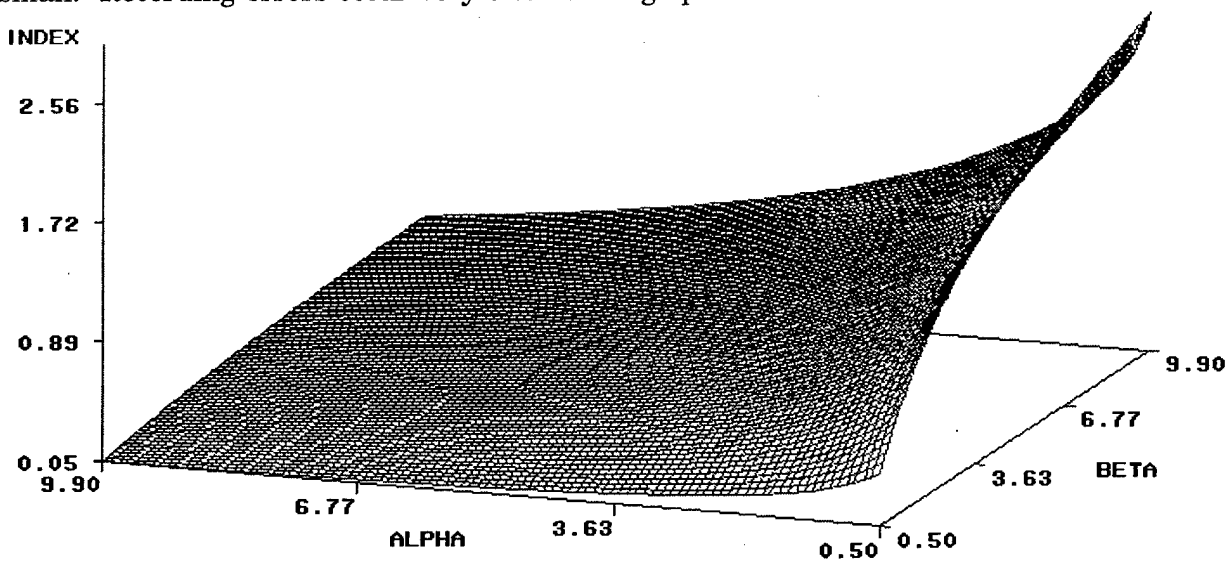



**Figure 3. The expected Shannon index with alpha and beta ranged from 0.5 to 10.**

as ABCD1 and ABCD2. When there are many species with similar species codes in one plot, the recording errors can be very large, contributing the most to the total misclassification. Miscoding species happens usually at data entry. Incorrectly typing a code results in a "new species." This error can alter the Shannon index significantly. Specimens of unknown species are often collected from the field for further identification. If the specimens are not handled properly, or are stored for too long, accurate identification becomes difficult.

If only misclassification is considered, the error in absolute species abundance is linear in relation to the error in relative species abundance. This is because the population is a closed system when only misclassification error is considered. That is to say, misclassification does not change the total population; it only changes the species distributions. The error for the total population due to misclassification is zeros. Therefore, error in relative species abundance is linear to the error in absolute species abundance.

Let $p_s$ be the relative abundance of species s without misclassification, and $p_s'$ be the relative abundance of species s with misclassification and $e_s$ be the misclassification error in relative species abundance. We have

$$p_s' = p_s + e_s \tag{33}$$

where $p_s$ is beta distributed. Because the population is a closed system, $p_s'$ must be also beta distributed. This property makes it difficult to find a distribution for the error term $e_s$. Misclassification actually just shifts the beta distribution from $L(p_s \mid \alpha, \beta)$ to $L(p_s' \mid \alpha', \beta')$, where $\alpha$ and $\beta$ are the parameters of beta function without misclassification, and $\alpha'$ and $\beta'$ are the parameters of beta function with misclassification. Instead of finding the distribution for the error term $e_s$, we look for the relationship between the error and distribution shift. For example, when we say 20% misclassification error in species abundance, we imply that the difference of area between the $L(p_s \mid \alpha, \beta)$ and $L(p_s' \mid \alpha', \beta')$ curves is 20%. In general, we have

$$\frac{\int_0^1 \mid L(p_s \mid \alpha, \beta) - L(p_s' \mid \alpha', \beta') \mid p_s dp_s}{\int_0^1 p_s L(p_s \mid \alpha, \beta) dp_s} = d\% \tag{34}$$

where d is the misclassification error as a percentage. Thus, the conditional expectation of Shannon index $E(H \mid d)$ is calculated from the sample curves of $(\alpha', \beta' \mid d)$. The bias and variance due to misclassification are given by

$$(35) \quad bias(H) = E(H|\alpha,\beta) - E(H|d)$$

$$Var(H) = Var(H|\alpha,\beta) + Var(H|d)$$

where $E(H \mid \alpha, \beta)$ and $E(H \mid d)$ are the expectation of the Shannon index without misclassification and with d% input misclassification, respectively. Similarly for variance $Var(H \mid \alpha, \beta)$ and $Var(H \mid d)$.

A C program was written to model the errors of the Shannon index due to misclassification. The program ran on a LENA SUPERCOMPUTER of the National Center for Supercomputing Applications (NCSA) in Champaign, Illinois. In the simulation, we took ($\alpha$, $\beta$) in the likelihood function as (1.0, 1.0), (2.0, 2.0), (3.0, 0.5), and (0.5, 3.0) for flat *priori*, bell-shaped, J-shaped, and inverse J-shaped curves respectively. The results from the program are used to generate the tables discussed in the following paragraphs.

Tables 1 through 4 show the bias and variance of the Shannon index for the four typical types of species abundance distributions. In Table 1 the estimation of the Shannon index is calculated based on Equations (31) and (32), which is the case without misclassification. The error in Table 1 is mainly due to the natural variation of species distribution, which can be used to determine sampling design. The bias and error in Tables 2 through 4 only count for the misclassification of two input error limits of 10% and 20%. Comparing Tables 1 and 2, we can see that the random misclassification does not contribute much to the variance, but it does produce bias even if the input data is unbiased.

Table 2 shows the bias and variance of the Shannon index due to random misclassification. From this table we can see that different species distributions have different sensitivities to misclassifications. The flat *priori* and bell shaped distributions have the most resistance to misclassifications. With 10% input misclassification, the bias is about 1.75%, part of which may due to rounding error in the computation of the complete beta function $B$ ($\alpha$, $\beta$). With larger input error (20%), misclassification causes about 6.77% bias in the Shannon index. One of the reasons for the lower sensitivity of misclassification in the flat *priori* and bell shaped distributions is that the chances of incorrectly classifying rare species as common species and vice versa are relatively equal. In other words, the effects of misclassification are canceled out by each other. In the J-shaped case, the chances of incorrectly classifying rare species as common species are less than that of incorrectly classifying common species as rare species because

there are a few rare species in a community dominated by common species. Misclassification may create "new species" or increase the evenness of species distribution. This causes systematic increase of the Shannon index. With 10% input misclassification, the Shannon index increases about 5%. With 20% misclassified, the Shannon index increases 11%. In contrast with the inverse J-shaped case, the Shannon index decreases with misclassification. The Shannon index decreases 4% with 10% input error and 18% with 20% input error. In the inverse J-shaped case, because there are many rare species and very few dominating species, it is more likely to overlook some rare species. This will reduce the number of species and causes the Shannon index to decrease.

Tables 3 and 4 give the bias and variance of the Shannon index due to systematic misclassification. As an example, we suppose the bias of input error is linear to its Shannon index, $E(H \mid d) = E(H \mid \alpha, \beta) + w * E(H \mid d)$. As shown in Tables 3 and 4, both the bias and variance of the Shannon index are larger with systematic misclassification. Table 3 illustrates that more common species are misclassified as rare species or species are misidentified as "new species." That is, the misclassification increases the evenness of species distribution or number of species. This causes the Shannon index to increase. Table 4 demonstrates that more rare species are misclassified as common species. This drops the evenness of species distribution and causes a decrease in the Shannon index.

**Table 1. Shannon index without misclassification.**

| Model | flat priori | bell-shaped | J-shaped | Inverse J-shaped |
|---|---|---|---|---|
| mean | 0.5 | 0.5833 | 0.1388 | 1.4132 |
| variance | 0.0461 | 0.0257 | 0.0165 | 0.0723 |
| Error[1] (%) | 42.8945 | 27.4938 | 92.59 | 60.1770 |

[1] Error (%) = (standard deviation)/(true index)* 100

**Table 2. Bias and variance of Shannon index due to random misclassification.**

| Model | Flat priori | | Bell-shaped | | J-shaped | | Inverse J-shaped | |
|---|---|---|---|---|---|---|---|---|
| Input error | 10 % | 20 % | 10 % | 20 % | 10 % | 20 % | 10 % | 20 % |
| Mean | 0.5088 | 0.5338 | 0.5885 | 0.608 | 0.1458 | 0.1553 | 1.347 | 1.1529 |
| Variance | 0.0022 | 0.0095 | 0.002 | 0.0095 | 0.0001 | 0.0003 | 0.0711 | 0.0818 |
| Error (%) | 9.4453 | 19.5184 | 7.7545 | 16.6964 | 6.5929 | 12.4139 | 18.8735 | 20.2347 |
| Bias | 0.0088 | 0.0338 | 0.0052 | 0.0247 | 0.007 | 0.0165 | -0.0662 | -0.2603 |
| Bias (%) | 1.7573 | 6.7682 | 0.8895 | 4.2423 | 5.0707 | 11.9189 | -4.6878 | -18.4224 |

**Table 3. Bias and variance of Shannon index due to weighted misclassification of common species as rare species.**

| Model | Flat priori | | Bell-shaped | | J-shaped | | Inverse J-shaped | |
|---|---|---|---|---|---|---|---|---|
| Input error | 10 % | 20 % | 10 % | 20 % | 10 % | 20 % | 10 % | 20 % |
| Mean | 0.5597 | 0.6406 | 0.6473 | 0.7297 | 0.1604 | 0.1864 | 1.4816 | 1.3834 |
| Variance | 0.0027 | 0.0137 | 0.0025 | 0.0137 | 0.0001 | 0.0004 | 0.0861 | 0.1197 |
| Error (%) | 10.3899 | 23.4221 | 8.53 | 20.0357 | 7.2522 | 15.0208 | 20.7608 | 24.484 |
| Bias | 0.0597 | 0.1406 | 0.064 | 0.1464 | 0.0216 | 0.0476 | 0.0684 | -0.0298 |
| Bias (%) | 11.933 | 28.1219 | 10.9784 | 25.0907 | 15.5778 | 34.3027 | 4.8434 | -2.1069 |

**Table 4. Bias and variance of Shannon index due to weighted misclassification of rare species as common species.**

| Model | Flat priori | | Bell-shaped | | J-shaped | | Inverse J-shaped | |
|---|---|---|---|---|---|---|---|---|
| Input error | 10 % | 20 % | 10 % | 20 % | 10 % | 20 % | 10 % | 20 % |
| Mean | 0.463 | 0.4484 | 0.5355 | 0.5108 | 0.1327 | 0.1305 | 1.2257 | 0.9684 |
| Variance | 0.0018 | 0.0067 | 0.0017 | 0.0067 | 0.0001 | 0.0002 | 0.0589 | 0.0577 |
| Error (%) | 8.5953 | 16.3955 | 7.0566 | 14.025 | 5.9996 | 10.4277 | 17.1749 | 16.9972 |
| Bias | -0.037 | -0.0516 | -0.0478 | -0.0725 | -0.0061 | -0.0083 | -0.1875 | -0.4448 |
| Bias (%) | -7.4009 | -10.3147 | -8.1906 | -12.4365 | -4.3857 | -5.9881 | -13.2659 | -31.4748 |

### *Bayesian estimation method*

A basic assumption of the method of likelihood function is that species abundance distribution follows a beta distribution. The distribution of real-world species abundance may not follow any theoretic distribution. To adjust for the misclassification in such a case, misclassification probability or a double sampling scheme is used. The following methods are adopted from Viana (1994) and Geng (1989).

Let us consider binomial data first. The extension from binomial to multinomial data is straightforward. Let x = ($x_1$, $x_2$) be the observed binomial data subject to misclassification. Let $p'$ = ($p_1'$, $p_2'$) and p = ($p_1$, $p_2$) be the corresponding observed and true probability distributions, respectively, and M be the 2 x 2 matrix of classification error probabilities, so that $p' = M\,p$, or

$$(36) \qquad \begin{bmatrix} p_1' \\ p_2' \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \end{bmatrix}$$

where $m_{ij}$ is the conditional probability of observed state j given the true state i. There are two cases based on the knowledge of the misclassification probability matrix M: M is known and M is unknown. We will discuss two methods corresponding to these two cases.

**Cases where M is known.** When M is known, estimates of true classification probability p are obtained from

$$(37) \quad p = M^{-1}p'$$

under the condition that $M^{-1}$ exists. In practice, we have matrix M and data x = $(x_1, x_2)$. We wish to know the expectation and variance of the estimated p from M and x. Viana (1994) gave the posterior density of p, given x and M.

The posterior density $f(p \mid x, M)$ is a weighted sum of beta density functions given by:

$$(38) \quad f(p|x, M) = \sum_{r \in \Re} \omega_r L(p; \alpha + \sum_u r_{u\sim}),$$

where $\omega_r = W_r / \sum_{i \in \Re} W_i,$

$$W_r = B(\alpha + \sum_u r_{u\sim}) \prod_u \left[ \binom{r_+}{r_{u\sim}} \prod_v m_{uv}^{r_{uv}} \right],$$

$$L(p;\alpha) = p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} / B(\alpha),$$
$$\mathrm{B}(\alpha) = \int_0^1 x^{a_1 - 1}(1-x)^{a_2 - 1} dx,$$

and $\Re$ is the set of all matrices r with entries $r_{uv}$ such that $r_{u+} = x_u$ with u=1 or 2.

If the observed data x, $x = (x_1, \ldots, x_k)$ is a multinomial vector, the posterior probability density $f(\pi \mid x, M)$ is a weighted sum of Dirichlet density functions (Viana 1994):

$$(39) \quad f(p|x, M) = \sum_{r \in \Re} \omega_r D(p; \alpha + \sum_u r_{u\sim}),$$

where $\omega_r$ and $W_r$ are the k-dimension of their definitions given above.

$$D(p; \alpha + \sum_u r_{u\sim})$$

is the Dirichlet function with the vector of parameters $(\alpha + \sum_u r_{u\sim})$.

**Cases where M is unknown.** When M is unknown, a double sampling scheme is used to calculate the posterior means of classification probabilities. When double

sampling, we observe individuals by using a cheap but error-prone method. We then categorize a random subsample by using a precise but expensive method to adjust for misclassification.

Let us consider a 2 x 2 table with two variables. In the discussion we will use notations given by Geng (1989):

A, B: error-free variables;

a, b: error-prone variables;

$\{m_{++jk}\}$: observed frequencies of main sample;

$\{n_{hijk}\}$: those of subsample;

$\{p_{hijk}\}$: cell probabilities;

$\{p_{hi|jk}\}$: conditional probabilities given a=j and b=k;

$\{\alpha_{hijk}\}$: parameters of Dirichlet density of $\{P_{hijk}\}$.

In these notations, '+' denotes a summation over the index. Indices h, i, j, and k denote variables A, B, a, and b, respectively. Suppose the observations are multinomial data and prior density of $\{P_{hijk}\}$ is a Dirichlet density with parameters $\{\alpha_{hijk}\}$,

$$(40) \quad D(\{p_{hijk}\}|\{\alpha_{hijk}\}) = \frac{\Gamma(\alpha_{hijk})}{\prod_{h,i,j,k}\Gamma(\alpha_{hijk})}\prod_{h,i,j,k} p_{hijk}^{\alpha_{hijk}-1}.$$

Geng (1989) gave the following the joint posterior density of $\{p_{++jk}\}$ and $\{p_{hi|jk}\}$, and the posterior means, variances and covariance of $\{p_{hijk}\}$:

$$(41) \quad f(\{p_{++jk}\},\{p_{hi|jk}\}|\{n_{hijk}\},\{m_{++jk}\}) = D(\{p_{++jk}\}|\{\alpha_{++jk}+n_{++jk}+m_{++jk}\})\cdot \prod_{j,k} D(\{p_{hi|jk}\}|\{\alpha_{hijk}+n_{hijk}\})$$

$$(42) \quad \dot{p}_{hijk} = \frac{(\alpha_{++jk}+n_{++jk}+m_{++jk})(\alpha_{hijk}+n_{hijk})}{(\alpha_{++++}+n_{++++}+m_{++++})(\alpha_{++jk}+n_{++jk})}$$

$$(43)\quad Var(p_{hijk}) = \frac{(\alpha_{++jk} + n_{++jk} + m_{++jk})(\alpha_{++jk} + n_{++jk} + m_{++jk} + 1)}{(\alpha_{++++} + n_{++++} + m_{++++})(\alpha_{++++} + n_{++++} + m_{++++} + 1)}$$

$$\frac{(\alpha_{hijk} + n_{hijk} + 1)(\alpha_{hijk} + n_{hijk})}{(\alpha_{++jk} + n_{++jk} + 1)(\alpha_{++jk} + n_{++jk})} - (p'_{hijk})^2$$

## Error Budget and Sensitivity Analysis

Suppose we have a discrete stochastic system without input control

$$(44)\quad \begin{aligned} x_{k+1} &= A_k x_k + G_k w_k \\ y_k &= C_k + H_k v_k, \end{aligned}$$

where $x_k \in R^n, y_k \in R^p, w_k \in R^g, v_k \in R^h$; $A_k, G_k, C_k$, and $H_k$ are possible time-varying, known matrices of the appropriate dimension, x and y are respectively the state space and observation space. The basic random variables $\{x_0, w_0, .., v_0, ...\}$ are all independent and Gaussian with $x_0 \sim N(0, \sum_0), w_k \sim N(0, Q), v_k \sim N(0, R)$. The covariance are all known. The available information at time k is $z^k = y^k := (y_k, y_{k-1}, ...., y_0)$. The random variable $x_K$, $x_{K+1}$ and $y^K$ are jointly Gaussian. Denote

$$P_{k|k}(x_k | y^k) \sim N(x_{k|k}, \Sigma_{k|k}), and$$

$$P_{k+1|k}(x_{k+1} | y^k) \sim N(x_{k+1|k}, \Sigma_{k+1|k}).$$

$$\tilde{x}_{k+1|k} := x_{k+1} - x_{k+1|k},$$

$$y_{k|k-1} := E\{y_k \mid y^{k-1}\}, and$$

$$\tilde{x}_{k|k} := x_k - x_{k|k}.$$

$$\tilde{y}_{k|k-1} := y_k - y_{k|k-1}$$

The estimation model of the system is

$$(44)\quad x_{k+1|k+1} = A_k x_{k|k} + L_{k+1}[y_{k+1} - C_{k+1}A_k x_{k|k}],$$

$$x_{0|0} = L_0 y_0,$$

$$\Sigma_{k+1|k+1} = (I - L_{k+1}C_{k+1})\Sigma_{k+1|k},$$

$$\Sigma_{k+1|k} = A_k \Sigma_{k|k} A_k^T + G_k Q G_k^T.$$

$$S_{0|0} = (I - L_0 C_0)S_0,$$

where $L_k = \Sigma_{k|k-1} C_k^T [C_k \Sigma_{k|k-1} C_k^T + H_k R H_k^T]^{-1},$

$$L_0 = \Sigma_0 C_0^T [C_0 \Sigma_0 C_0^T + H_0 R H_0^T]^{-1}$$

The n-step prediction model of the system is

$$(45)\quad x_{k+n|k} = \prod_{i=0}^{n-1} A_{k+i} x_{k|k},$$

$$(46)\quad \Sigma_{k+n|k} = \prod_{i=0}^{n-1} A_{k+i} \Sigma_{k|k} (\prod_{i=0}^{n-1} A_{k+i})^T + \sum_{i=0}^{n-1} G_{k+i} Q G_{K+i}^T$$

If $x_k$ represents the vector of species abundance at time k, the Shannon index can be calculated from the vector of species abundance $x_k$.

Let us rewrite the formula of Shannon index.

$$(47)\quad H = -\sum_{i=1}^{s} p_i \times \log(p_i)$$

where $p_i = x_{ik} / \sum_{j=1}^{s} x_{jk}$ and s is the number of species. Because H is a nonlinear function of $x_K$ we can calculate its error by the Taylor series expansion method as shown below.

Assuming an exact function **f** is used to make predictions:

(48) **Y=f(B,X)**

where **Y** is a prediction made with the function, $\mathbf{X}=(X_1, X_2, \ldots, X_m)$ is a vector of input variables, and $\mathbf{B}=(b_1, b_2, \ldots, b_m)$ is a vector of known parameters. **X** is usually assumed to be error-free. Now suppose instead of being error-free, the j-th component of **X**, $X_j$, has random error $e_j$ (i.e., $x_j = x_j + e_j$), where $e_j$s are independently distributed with mean 0 and variance $V(e_j)$. Then, the predicted **Y** also has error due to the errors of **X**. This error can be estimated by Taylor series expansion.

Denote $f(x) = -x \times \log(x)$, the first order and second order derivatives of $f(x)$ are

$$(49) \quad \frac{df(x)}{dx} = -\log(x) - 1$$

$$(50) \quad \frac{d^2 f(x)}{d^2 x} = -\frac{1}{x}$$

Applying (49) and (50) in the error propagation models (3) and (4), we have the expectation and variance of the Shannon index

$$(51) \quad E(H) = -\sum_{i=1}^{s} p_i \times \log(p_i) + \sum_{i=0}^{s} E(e_i) * (-\log(p_i) - 1) - \frac{1}{2}\sum_{i=0}^{s} E[e_i^2 * \frac{1}{p_i}]$$

$$(52) \quad Var(H) \approx \sum_{i=0}^{s} Var(e_i) * (-\log(p_i) - 1)^2$$

where $p_i = x_{i,k|k} / \sum_{j=1}^{s} x_{j,k|k}$, $e_i$ is the error of $p_i$.

### *Common inventory errors in belt transect*

Since 1989, the U.S. Army has delineated permanent core plots on over 50 military installations and training areas in the United States and Germany. The standard size of an LCTA permanent plot is 100 x 6 meters ($600m^2$) with a 100-m line transect forming the longitudinal axis. The plot inventory is conducted over a 2- to 3-month period during the peak of the growing season. The inventory consists of four major elements; land use assessment, line transect, belt transect,

and wildlife sampling. The belt transect is intended to characterize species composition, density, and height distribution of woody and succulent vegetation.

The belt transect extends the width of the 100-m line transect. Although the belt has a standard size of 6m, the width may be reduced at the field crews discretion for high-density species. The field data is subsequently extrapolated to a standard 100 x $6m^2$ plot during data analysis. Subject matter experts have suggested the following sources of inventory errors associated with LCTA belt transect methods:

1. Instrument error. The major instrument error is from locating the pole and tape positions. The tape may not go straight from one point to the other due to the dense plants. The tape may represent different paths for different years of data collection.
2. Observer's error. This error refers to the differences in inventory results made by different observers on the same plot. One of the major differences is in the way each person counts clumps. Clumps are dense clonal patches of individual stems. Some observers count a clump as a single plant, while others may count each branch of a clump as a separate plant.
3. Recording error. A common mistake is made when recording between species with similar codes such as ABCD1 and ABCD2. Recording errors result from different species having very similar codes. Recording errors also occur because codes are often truncated to help reduce plot measurement times.
4. Species recognition error. Species are misidentified due to poor quality of the specimen and when the specimen has characteristics that are similar to other species. Species recognition errors can result from field crews with varying levels of training and experience.
5. Expansion factor error. Data from reduced belt transects must be extrapolated to a standard size plot of 100 x $6m^2$. Even for the same species in the same plot, different observers may use different belt widths. The data may also be expanded by the wrong factor due to the changes of the expansion factor and a lost record of the expansion factor.
6. Editing error. Editing error refers to errors made in data entry. A common mistake is made when entering the wrong species code into the database. Entering the wrong species means creating a "new species."

A summary of the subject matter experts characterization of LCTA belt transect inventory error sources is provided in Table 5.

**Table 5. Suggested inventory error limits for belt transect.**

| Error sources | Lower bound | Upper bound |
|---|---|---|
| Locating poles and tapes | 10 | 30 |
| Counting clumps | 20 | 40 |
| Recording error | 0 | 5 |
| Species recognition | 5 | 10 |
| Temporal shift | 0 | 40 |
| Editing error | 0 | 30 |
| Expansion factor | 0 | 200 |

### *An example of error budget analysis*

We chose to use plant community type 6 from White Sands Missile Range data set (Cao et al. 2000) as an example of error-budget analysis. Community type 6 covers plots 21, 22, 64, 138, 160, 164, and 167. The species distribution of this plant community type is similar to an inverse J-shaped beta distribution with alpha 0.5 and beta 3.0. We ran the error-budget model with two types of error limits: small input error and large input error. The results are shown in Tables 6 and 7. Percent bias, percent error, and total error are calculated as:

$$\text{Bias } \% = \frac{\text{index with error - true index}}{\text{true index}} \times 100\%$$

$$\text{Error } \% = \frac{\text{standard deviation}}{\text{true index}} \times 100\%$$

$$\text{Total error} = \sqrt{\sum_i e_i}$$

Tables 6 and 7 show the error of the Shannon index with small and large input errors for plant community type 6 at White Sands Missile Range. For the current estimate of the Shannon index, the major error is from misclassification. Misclassification is mainly due to the misidentification of species, recording species in the wrong code, and mistyping species codes into the data set. The major measurement error is due to the method of estimating number of stems in clumps of plants. Making more consistent measurements in clumps and quality control of data entry can largely reduce these types of errors. For the 10-year prediction, the system error and modeling error account for a larger component of the total error. With a good understanding of the system and system model, the system error usually is small. When more data are collected through contin-

ued monitoring, the modeling error can be reduced and ignored if the data set is large enough.

**Table 6. Error-budget table of Shannon index with small input errors.**

| Error sources | Input Error % | Current Estimate | | 10-year Prediction | |
|---|---|---|---|---|---|
| | | Bias % | Error % | Bias % | Error % |
| System error | 5 | 0 | 1.49 | 0 | 14.86 |
| Sampling | 10 | 0 | 2.97 | 0 | 2.97 |
| Modeling | 5 | 0 | 1.49 | 0 | 14.86 |
| Measurement | 10 | 0 | 2.97 | 0 | 2.97 |
| Misclassify | 10 | -4.69 | 18.87 | -4.69 | 18.87 |
| Total | 18.7 | -4.69 | 19.45 | -4.69 | 28.85 |

**Table 7. Error-budget table of Shannon index with large input errors.**

| Error Sources | Input Error % | Current Estimate | | 10-year Prediction | |
|---|---|---|---|---|---|
| | | Bias % | Error % | Bias % | Error % |
| System error | 10 | 0 | 2.97 | 0 | 29.72 |
| Sampling | 50 | 0 | 14.86 | 0 | 14.86 |
| Modeling | 10 | 0 | 2.97 | 0 | 29.72 |
| Measurement | 20 | 0 | 5.94 | 0 | 5.94 |
| Misclassify | 20 | -18.42 | 20.23 | -18.42 | 20.23 |
| Total | 59.16 | -18.42 | 26.13 | -18.42 | 49.31 |

# 3 Conclusions

The benefits of an error-budget model can be substantial. First, the error-budget model evaluates the statements made from the survey. Given all errors, the error-budget model can determine if the statements are valid. A statement is valid only if it is within certain error limits. The statement provides little useful information if its errors are out of the specified limits. Second, the error-budget model guides survey decisions. With error sensitivity analysis, all types of error sources can be tested to find their effect on the final statement. Effort can be put into survey effort that controls the sources of error. In this manner, maximum accuracy can be obtained with minimum cost. Third, the error-budget model provides information on error correction. To correct errors, the sources of errors must first be known. Errors from different sources may require different correction procedures. Using error decomposition, the major causes of errors can be determined.

Error-budget analysis of the plant population model yielded a number of possible sources of error. For the initial estimates of the Shannon index, the major error was from misclassification. Misclassification is mainly due to misidentified species, recording species with the wrong code, and mistyping species codes into the data set. The major measurement error stemmed from the way clumps were counted. Over the course of a 10-year prediction, the system error and modeling error compounded, causing a significant rise in the total error.

Uncertainty in near term model predictions was largely determined by errors associated with data collection. These types of errors can be largely reduced by making consistent measurements and with an effective quality assurance/control program. Costs associated with reducing these sources of errors should be minimal. However, as model prediction periods increase, modeling and system errors become the most important source of uncertainty. These sources of error can be reduced only through an intimate knowledge of the ecology and model. As more data are collected through the years, the potential exists for modeling error to be reduced.

The error-budget model presented in this report illustrates the potential of using error budgets to assist land managers. The error budget provides the user of the model with a means to assess management alternatives. The consequences of alternative data collection and quality control procedures on model predictions

can be objectively assessed. Depending on the time period of concern, the error budget identifies sources of error that most affect decision making processes.

Based on the results of this study, it is recommended that uncertainty analysis tools such as error budgets be used more frequently in natural resources modeling efforts. Through the use of error budgets, interpretation of model results and subsequent management decisions can more accurately reflect our real understanding of the managed resources.

# References

Belcher, D., M. Holdaway, and G. Brand. 1982. *A description of STEMS, the stand and tree evaluation and modeling system.* Gen. Tech. Rep. NC-79, U.S. Forest Service.

Cao, Xiangchi, George Z. Gertner, Alan B. Anderson, and Bruce A. MacAllister. 2000. *Stochastic Model of Plant Diversity: Application to White Sands Missile Range.* Engineer Research and Development Center/Construction Engineering Research Laboratory (ERDC/CERL) Technical Report 00-5/ADA374140, February 2000.

Chen, T.T. 1979. "Log-linear models for categorical data with misclassification and double sampling." *J. Am. Statist. Assoc.*, 74, 481-488.

Chen, T.T. 1989. "A review of methods for misclassified categorical data in epidemiology." *Statist. Medicine*, 8, 1095-1106.

Diersing, V.E., R.B. Shaw, and D.J. Tazik. 1992. "US Army Land Condition-Trend Analysis (LCTA) Program." *Environmental Management* 16:405-414.

Gelb, A., J. Kasper, Jr., R. Nash, Jr., C. Price, and A. Sutherland, Jr. 1974. *Applied optimal estimation.* The MIT Press, Cambridge, MA. 374 pp.

Geng, Z. 1989. "Bayesian Estimation Methods for Categorical Data With Misclassifications." *Commun. Statist. -Theory Meth.*, 18(8), pp 2935-2954.

Gertner, G.Z. 1987. "Approximating precision in simulation projections: an efficient alternative to Monte Carlo methods." *Forest Service* 33: 230-239.

Gertner, G.Z. 1988. "Alternative methods for improving variance approximation of single tree growth and yield projections." In: A.R. Ed, S.R. Shifley, and T.E. Burk (Editors), *Proc. Forest Growth Modelling and Prediction, Vol II*, Gen. Tech, Rep. NC-120, North Central Forest Experiment Station, USDA Forest Service, pp 739-746.

Gertner, G.Z. 1990a. "Error budgets: A means of assessing component variability and identifying efficient ways to improve model predictive ability." In: R. Dixon, R. Meldahl, G. Ruark, and W. Warren (Editors), *Forest Growth Process Modeling of Responses to Environmental Stress.* Timber Press, Portland, OR, p 220.

Gertner, G.Z. 1990b. "The sensitivity of measurement error in stand volume estimation." *Can. J. For. Res.*, 20:800-804.

Gertner, G.Z. 1991. "Prediction bias and response surface curvature." *For. Sci.*, 37(3):755-765.

Gertner, G.Z., and M. Köhl. 1992. "An assessment of some nonsampling errors in a national survey using an error budget." *For. Sci.*, 38(3):525-538.

Gertner, G.Z., X. Cao, and H. Zhu. 1995. "A Quality Assessment of a Weibull Based Growth Projection System." *Forest Ecology and Management*, 71:235-250.

Gotelli, N. 1998. *A primer of ecology*. Sinauer Associates, Sunderland, MA.

Magnussen, S., and T.J.B. Boyle. 1995. "Estimating sample size for inference about the Shannon-Weaver and the Simpson indices of species diversity." *Forest Ecology and Management*, 78:71-84.

Mowrer, H. 1988. "A Monte Carlo comparison of propagated error for two types of growth models." In: A.R. Ed, S.R. Shifley, and T.E. Burk (Editors), *Proc. Forest Growth Modelling and Prediction, Vol II*. Gen. Tech. Rep. NC-120, North Central Forest Experiment Station, USDA Forest Service, pp 778-785.

Mowrer, H., and W. Frayer. 1986. "Variance propagation in growth and yield projections." *Can. J. For. Res.*, 16:1196-1200.

Shannon, C.E., and W. Weaver. 1949. *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, IL.

Tenebein, A. 1979. "A double sampling scheme for estimating binomial data with misclassification." *J. Am. Statist. Assoc.*, 65, 1350-1361.

Viana, M. 1994. "Bayesian Small-sample Estimation of Misclassified Multinomial Data." *Biometrics*, 50, pp 237-243.

York, J.C. 1992. "Bayesian methods for the analysis of misclassified or incomplete multivariate discrete data." Ph.D. dissertation, Department of Statistics, University of Washington, Seattle, WA.

## Distribution

Chief of Engineers
ATTN: CEHEC-IM-LH (2)
ATTN: HECSA Mailroom (2)
ATTN: CECC-R
ATTN: CERD-L
ATTN: CERD-M

SERDP (5)

ACS(IM)
ATTN: DAIM-ED-N (2)

Defense Tech Info Center 22304
ATTN: DTIC-O (2)

16
11/96

# REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188



| 1. AGENCY USE ONLY (Leave Blank) | 2. REPORT DATE | 3. REPORT TYPE AND DATES COVERED |
|---|---|---|
| | April 2000 | Final |

**4. TITLE AND SUBTITLE**
Errors in Environmental Assessments: An Error-Budget Model for Plant Populations

**5. FUNDING NUMBERS**
EE9
62720
A896
CN-T09
SERDP CS-1096

**6. AUTHOR(S)**
Xiangchi Cao, George Gertner, Bruce MacAllister, and Alan Anderson

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**
U.S. Army Engineer Research and Development Center
Construction Engineering Research Laboratory (ERDC/CERL)
P.O. Box 9005
Champaign, IL 61826-9005

**8. PEFORMING ORGANIZATION REPORT NUMBER**
ERDC/CERL TR-00-12

**9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
Strategic Environmental Research and Development Program
ATTN: SERDP
901 N Stuart St., Suite 303
Arlington, VA 22203-1853

Headquarters, Department of the Army
ATTN: DAIM-ED-N
Assist Chief of Staff (Installation Mgmt)
600 Army Pentagon
Washington, DC 20310-0600

**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**

**9. SUPPLEMENTARY NOTES**
Copies are available from the National Technical Information Service, 5385 Port Royal Road, Springfield, VA 22161

**12a. DISTRIBUTION / AVAILABILITY STATEMENT**
Approved for public release; distribution is unlimited.

**12b. DISTRIBUTION CODE**

**13. ABSTRACT** *(Maximum 200 words)*

The assumption that a data set used for any mathematical or computer modeling is error-free is an underlying premise of theoretical modeling. However, the assumptions of error-free data and models usually do not hold true in the real world. Error is a natural property of surveys and modeling. Consequently, error should be taken into account when developing any type of model. The goal of this project is to create an error-budget model for a population dynamics model of plant communities. Once developed, this error-budget model can in turn be used for a number of other purposes such as data correction, model evaluation, quality control, and management decisionmaking.

**14. SUBJECT TERMS**
environmental assessment
Land Condition Trends Analysis (LCTA)
Strategic Environmental Research and Development Program (SERDP)
plant communities
data management
natural resources management
modeling

**15. NUMBER OF PAGES**
42

**16. PRICE CODE**

| 17. SECURITY CLASSIFICATION OF REPORT | 18. SECURITY CLASSIFICATION OF THIS PAGE | 19. SECURITY CLASSIFICATION OF ABSTRACT | 20. LIMITATION OF ABSTRACT |
|---|---|---|---|
| Unclassified | Unclassified | Unclassified | SAR |